基于高效门控与目标区域关注的实时视频超分辨率

林乐平[1,2]，苏 治[2]，欧阳宁[1,2]

（1.桂林电子科技大学 广西无线宽带通信与信号处理重点实验室，广西 桂林 541004；2.桂林电子科技大学 信息与通信学院，广西 桂林 541004）

摘 要：面对大运动幅度的复杂视频场景，实时视频超分辨率算法难以重建纹理细节、遮挡区域。本文基于生成对抗网络，提出了一种基于高效门控与目标区域关注的实时视频超分辨率方法。该方法首先使用高效门控重建网络作为生成网络，在保持高效的同时，通过简化的门控机制自适应选择复杂区域信息，以提升重建结果。进一步地，该方法提出了目标区域关注鉴别网络，为生成网络提供多尺度及时空信息反馈，通过多尺度机制和ReLU线性注意力获取复杂视频的多尺度信息；通过显著时空鉴别模块，限制鉴别网络关注复杂区域，以更好地获取复杂区域的时空信息。实验结果表明，所提方法相较于其他先进算法具有显著的优越性；在模型效率方面，实现了13.36 ms的推理延迟及65.806的实时得分，显示了模型高效的实时性能。

关键词：视频超分辨率；实时；生成对抗网络；门控机制；目标区域关注

中图分类号：TP391.4 **文献标志码**：A **文章编号**：1671-5497(2026)03-0819-11

DOI：10.13229/j.cnki.jdxbgxb.20240926

Efficient gating and target region attention based real-time video super-resolution

LIN Le-ping[1,2], SU Zhi[2], OUYANG Ning[1,2]

(1. *Guangxi Key Laboratory of Wireless Broadband Communication and Signal Processing, Guilin University of Electronic Technology, Guilin* 541004, *China*; 2. *School of Information and Communication, Guilin University of Electronic Technology, Guilin* 541004, *China*)

Abstract: Facing complex video scenes with large motion amplitude, real-time video super-resolution algorithms are difficult to reconstruct texture details and occluded regions, based on generative adversarial network, a real-time video super-resolution method based on efficient gating and target region attention is proposed. The method firstly uses an efficient gating reconstruction network as a generative network to maintain high efficiency while adaptively selecting complex region information through a simplified gating mechanism to enhance the reconstruction results. Further, the method proposes a target region attention

收稿日期：2024-08-23.

基金项目：国家自然科学基金项目(62001133)；广西科技基地和人才专项项目(桂科AD19110060)；广西自然科学基金项目(2017GXNSFBA198212)；广西无线宽带通信与信号处理重点实验室基金项目(GXKL06200114).

作者简介：林乐平(1980-)，女，教授，博士.研究方向：机器学习，图像信号处理.E-mail：linleping@guet.edu.cn

通信作者：欧阳宁(1972-)，男，教授，硕士.研究方向：数字图像处理，智能信息处理.E-mail：ynou@guet.edu.cn

discriminative network to provide multiscale and spatio-temporal information feedback for the generative network, and acquires multiscale information of the complex video through the multiscale mechanism and ReLU linear attention; through the significant spatio-temporal discriminative module, it restricts the discriminative network to focus on the complex regions, and better acquires the spatio-temporal information of the complex regions. The experimental results show that the proposed method exhibits significant superiority over other SOTA algorithms, the model efficiency achieves an inference delay of 13.36 ms and a real-time score of 65.806, which indicates the efficient real-time performance of the model.
Key words: video super-resolution; real-time; generative adversarial network; gating mechanisms; target region attention

0 引 言

视频超分辨率(Video super resolution,VSR)是计算机视觉的核心问题之一,其目标是从给定的低分辨率(Low resolution,LR)视频序列中重建出高分辨率(High resolution,HR)视频序列。然而,在大运动幅度的复杂视频场景,如交通车流、体育赛事、电影场景等,存在大量纹理细节及遮挡区域,VSR模型难以获取这些视频场景的特征信息,无法生成高质量视频,这是VSR所面临的挑战之一。

随着深度学习的兴起,基于深度学习的模型在VSR领域取得了可观的成就。对于VSR的深入研究表明,帧间信息的利用直接影响到VSR的性能[1]。有效、充分地利用帧间信息可提高VSR质量,包括利用光流进行运动估计对齐相邻帧和特征。Caballero等[2]使用从细到粗的光流与图像超分辨率(Super resolution,SR)模型结合,提出了实时VSR模型。Vemulapalli等[3]将光流运动估计与先前推断的HR估计扭曲。Wang等[4]直接从LR帧估计HR光流。Chan等[5]通过光流传播邻近特征,但是在处理复杂、运动幅度大的视频场景时、光流估计易产生误差。为解决这个问题,Chan等[6]将可变形卷积与几何建模变换结合,更有效地利用视频帧的帧间信息。Ouyang等[7]引入门控机制,自适应选择视频帧信息,以减少光流估计误差的影响。Zhou等[8]利用相邻帧的特征级时间连续性来减少冗余计算,更合理地利用先前增强的SR特征。VSR技术在复杂视频场景中的应用对于提升视频质量、优化资源利用、提高视频内容等方面具有重要意义。

如今在线会议、视频直播的日益普及,实时视频增强越来越受到研究者的关注。实时VSR成为充满挑战性和重要意义的VSR领域研究问题。Cao等[9]提出的EGVSR为一款高效且通用的实时VSR模型,其在提高推理速度的同时保持了模型性能。Xia等[10]使用结构化稀疏学习,对整体模型裁剪、轻量化的同时保持了显著竞争力的实时VSR模型性能。Xiao等[11]使用旁路卷积移植,通过蒸馏学习的方式提升实时VSR性能。如今的实时VSR大都研究实际部署问题,保持模型性能的同时降低推理延迟。但同时可以观察到,这些模型存在缺陷,首先,实时VSR模型的轻量级设计本质上减少网络纵深,通过保留少量卷积层来实现,这会导致模型无法充分进行特征提取;其次,为简化计算,实时VSR模型没有充分提取多尺度信息,对多尺度特征的融合过于简化,无法有效融合不同尺度的信息。而现实生活中的复杂视频场景存在丰富的多尺度信息,尤其是纹理细节及遮挡区域,这会导致实时VSR模型无法有效提取复杂区域的特征,严重限制模型性能;并且,实时VSR基本使用基于光流估计的时间聚合,在面对运动幅度大的复杂视频场景时易产生误差。这些问题导致实时VSR难以重建纹理细节、遮挡区域。Li等[12]通过保留复杂的网络结构,对数据进行时空过拟合,解决这些问题,但这会增加模型参数和推理延迟。最重要的是,与其他VSR研究相比,实时VSR对推理延迟的严格限制给重建视频复杂区域带来了更大的挑战,这对实时VSR模型效率至关重要。

为了提高视觉质量,生成对抗网络(Generative adversarial network,GAN)被应用于VSR领域,有利于生成更多细节纹理的HR帧。Chu等[13]提出一种对抗性和循环训练策略,将GAN用于VSR领域,该研究巧妙地平衡了空间高频细节还原与时间连续性之间的关系,实现实时VSR

的同时突破了视觉感知差的问题。Chen 等[14]则是多帧输入结合双鉴别器指导网络训练，为生成网络提供丰富的时空信息，指导生成网络更加关注像素特征。可以观察到，GAN 可通过训练的方式，在不增加生成网络计算复杂度的情况下，指导生成网络复杂视频场景时空信息，生成更多逼真的细节纹理的同时保持实时 VSR 的高效性能。

面对运动幅度大的复杂视频场景，针对实时 VSR 难以重建纹理细节、遮挡区域的问题，受到文献[15-18]的启发，本文基于 GAN，提出了一种基于高效门控与目标区域关注的实时 VSR 方法（Efficient gating and target region attention-based real-time VSR，EGTRA），该方法首先使用高效门控重建网络（Efficient Gating Reconstruction Network，EGRN）作为生成网络，利用运动估计及补偿进行时间聚合，获取相邻帧信息。其次，使用非线性门控模块（Nonlinear gating module，NGM），通过简易门控自适应选择视频复杂区域信息，减少时间聚合误差影响，重参数化残差块（Reparameterizable residual block，RRB）的深度特征提取，去除冗余滤波器，保证模型性能显著的同时减少模型参数，降低推理延迟。更进一步，该方法提出了目标区域关注鉴别网络（Target region attention discriminative network，TRADN），为生成网络提供多尺度及时空信息反馈，引导生成网络重建视频复杂区域，利用 ReLU 线性注意力（ReLU linear attention，RLA），通过多尺度记住关注复杂区域，获取复杂视频场景及复杂区域的多尺度信息，通过显著时空鉴别模块（Significant spatio-temporal discriminative module，SSTDM），限制鉴别网络关注视频复杂区域，获取复杂区域的时空信息。实验结果表明，相较于最新的轻量或实时 VSR 模型，本文所提方法能够有效提升实时 VSR 网络的性能，获得显著重建质量视频，在模型效率方面获得了极低的推理延迟及最高的实时得分，体现了模型高效的实时性能。

1　本文方法

为重建纹理细节和逼真纹理的连续重建视频帧，本文使用 GAN 框架构建了 EGTRA，其具有 1 个生成网络和 1 个鉴别网络，鉴别网络试图鉴别重建样本和真实样本。本文采用生成网络和鉴别网络之间的对抗性训练过程。生成网络如图 1 所示。

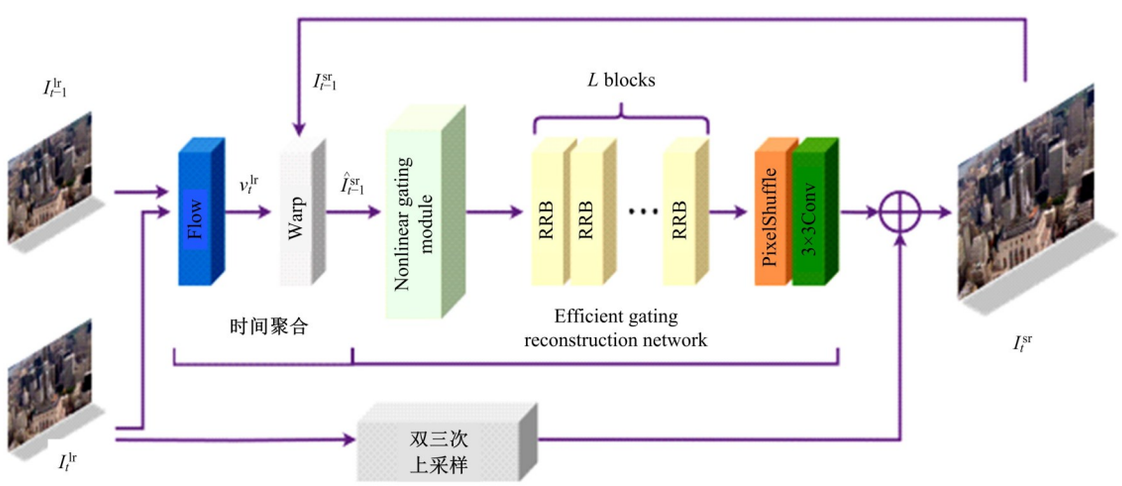

图 1　生成网络的整体框架示意图

Fig. 1　Illustration of the overall framework of generative network

生成网络由两部分组成，时间聚合和 EGRN。在时间聚合模块，通过光流对前一帧提取的特征进行对齐，对齐后的特征与上一重建帧提取的特征进行翘曲融合。在高效门控重建网络中，翘曲融合特征在 NGM 中提取视频复杂区域特征信息，输入重参数化残差块进行深度特征提取。生成网络的输出重建帧，是采用主路通过 Pixelshuffle 算子提取的特征进行上采样，与参考系双三次上采样相加，输出最终重建帧 I_t^{sr}。

1.1　时间聚合

对于时间聚合，本文采用 FRVSR[3]的光流网络，记为 $FNet(\cdot)$，估计 LR 帧 I_t^{lr} 与前一帧 I_{t-1}^{lr} 之间的运动估计，如式（1）所示：

$$v_t^{lr}=FNet(I_t^{lr},I_{t-1}^{lr}) \tag{1}$$

在运动补偿特征空间中,利用运动估计 v_t^{lr} 与前一帧 I_{t-1}^{sr} 重建帧进行对齐,获得的翘曲特征如式(2)所示:

$$\hat{I}_{t-1}^{sr}=Warp(v_t^{lr},I_{t-1}^{sr}) \tag{2}$$

式中:$\hat{I}_{t-1}^{sr}$ 为相应的翘曲特征;$Warp(\cdot)$ 为翘曲算子。

1.2 高效门控重建网络

在面对复杂视频场景时,尤其在大运动幅度下,对视频纹理细节及遮挡区域的时间聚合易产生误差,而门控机制可以有效缓解时间聚合带来的误差。门控机制凭借其高效的信息处理能力,在图像恢复领域已经取得广泛应用,为了更好地重建复杂视频场景下的视频细节特征,本文提出了EGRN,提升重建效果,其中,受Chen等[15]的启发,提出的NGM如图2所示。首先,对翘曲特征进行特征提取,得到特征 F_1,通过层归一化,得到特征 F_2,其过程如式(3)(4)所示:

$$F_1=Conv_{3\times3}(\hat{I}_{t-1}^{sr}) \tag{3}$$

$$F_2=LayerNorm(F_1) \tag{4}$$

式中:$Conv_{3\times3}(\cdot)$ 表示3×3卷积,$LayerNorm(\cdot)$ 表示层归一化。

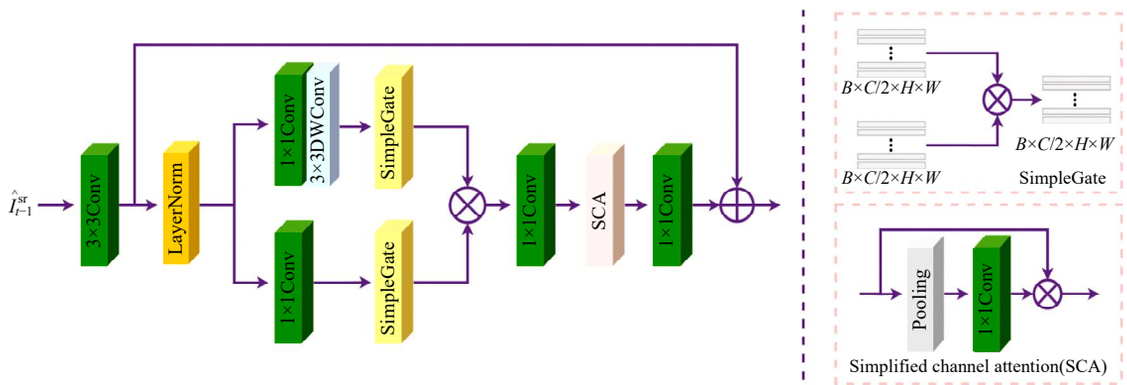

图2 NGM整体框架示意图

Fig. 2 Illustration of the overall framework of NGM

NGM主要由两条支路组成,特征 F_2 经过第一条支路,得到支路特征 F_a;经过第二条支路得到支路特征 F_b,两支路特征结合,得到特征 F_3,其过程如式(5)(6)(7)所示:

$$F_a=SG(Conv_{1\times1}(F_2)) \tag{5}$$

$$F_b=SG(DWConv_{3\times3}(Conv_{1\times1}(F_2))) \tag{6}$$

$$F_3=F_a\times F_b \tag{7}$$

式中:$SG(\cdot)$ 表示简易门控(Simple gate, SG),$Conv_{1\times1}(\cdot)$ 表示1×1卷积,$DWConv_{3\times3}(\cdot)$ 表示3×3膨胀卷积。

第一条支路通过小范围特征提取,去除时间聚合中的冗余信息,获得更精细的特征。第二条支路由1×1卷积和3×3膨胀卷积组成,串联起来,提取细粒度丰富的细节视频帧复杂区域特征及更为广泛的视频场景空间特征。之后引入非线性门控,即SG,SG在保留最少门控结构的同时,增强网络的泛化能力。采用门控机制调节信息流和处理抑制噪声,自适应聚焦复杂区域相关信息,提高网络特征提取能力。

由于特征 F_3 存在不同支路提取的特征,首先进行跨通道信息整合并减少通道数,得到特征 F_4,而后经过简化的通道注意力(Simplified channel attention,SCA),得到特征 F_5,与特征残差相加,得到NGM输出特征 F_{ngm},其过程如式(8)(9)(10)所示:

$$F_4=Conv_{1\times1}(F_3) \tag{8}$$

$$F_5=SCA(F_4)+F_4 \tag{9}$$

$$F_{ngm}=Conv_{1\times1}(F_5)+F_1 \tag{10}$$

式中:$SCA(\cdot)$ 表示SCA。

与通道注意力不同的是,SCA去除了非线性激活函数,只保留平均池化与卷积层,平均池化聚合全局信息,卷积层促进通道信息交互。

残差结构在神经网络中得到广泛应用,是必需的网络结构,但由于常规残差网络层较深,在残差连接时会导致额外内存消耗,由于额外占用内存降低了模型推理速度。为继承残差学习的优点且不产生额外内存占用,本文引入重参数化卷积[16],设计RRB如图3所示,RRB在训练阶段提取复杂区域的潜在特征,在推理阶段等效于一个3×3卷积,减少网络深度。

残差块特征提取过程如式(11)所示:

$$F_i=RRB(F_{ngm}),\ i=1,2,\cdots,L \tag{11}$$

式中:$RRB(\cdot)$ 表示RRB;F_i 表示经过不同层的

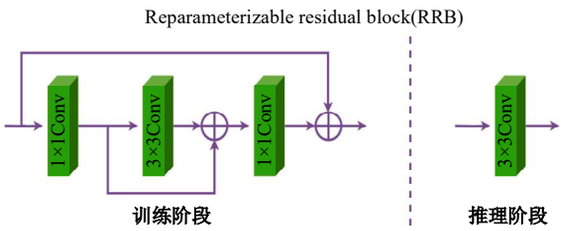

图 3　RRB 示意图

Fig. 3　Illustration of RRB

RRB 的特征输出。

最终残差输出 F_L 通过 Pixelshuffle 算子提取的特征进行上采样，与参考帧的双三次上采样结果相加，输出最终重建帧 I_t^{sr}。

1.3　目标区域关注鉴别网络

因复杂视频场景包含大量多尺度信息，对 VSR 而言，视频的多尺度与时空信息必不可少，为保持实时 VSR 模型高效，并使其能获取视频多尺度与时空信息，本文提出了 TRADN，其结构如图 4 所示。首先，输入真实样本三元组 $I^{\mathrm{gt}}=\{I_{t-1}^{\mathrm{gt}},I_t^{\mathrm{gt}},I_{t+1}^{\mathrm{gt}}\}$ 和重建样本三元组 $I^{\mathrm{sr}}=\{I_{t-1}^{\mathrm{sr}},I_t^{\mathrm{sr}},I_{t+1}^{\mathrm{sr}}\}$，鉴别网络可学习并理解视频帧的时空分布，受 Cai 等[17]的启发，通过线性投影层获得 Q/K/V，通过不同核大小的聚合卷积模块生成多尺度令牌，RLA 应用于多尺度支路，将输出连接并馈送至最终的线性投影层进行特征融合，得到包含复杂视频及复杂区域的多尺度信息输出 F_{input}，具体过程如式(12)～(15)所示：

$$F_6=RLA(Q,K,V) \tag{12}$$

$$F_7=RLA(GC_1(Q,K,V)) \tag{13}$$

$$F_8=RLA(GC_2(Q,K,V)) \tag{14}$$

$$F_{\mathrm{input}}=Linear(F_6,F_7,F_8) \tag{15}$$

式中：$Q=xW_Q$，$K=xW_K$，$V=xW_V$，W_Q、W_K、W_V 为可学习的线性投影矩阵，x 表示 I^{sr} 和 I^{gt} 拼接的输入；$RLA(\cdot)$ 表示 RLA；$GC_1(\cdot)$ 表示核大小为 3 的聚合卷积模块；$GC_2(\cdot)$ 表示核大小为 5 的聚合卷积模块；$Linear(\cdot)$ 表示线性投影层。

通过多尺度支路，可有效提取复杂区域的多尺度信息，同时保证足够的感受野提取复杂视频场景的多尺度信息。

随后，在 SSTDM 中，采用具有多尺度注意力机制的 EMA[18]，限制鉴别网络重点关注视频复杂区域，以获取视频时空信息，最终输出鉴别结果，其过程如式(16)所示：

$$P_{0/1}=DB^{(i)}(F_{\mathrm{input}})\quad i=1,2,3,4 \tag{16}$$

式中：$DB(\cdot)$ 表示显著时空鉴别块，其由卷积层、EMA、BatchNorm(BN)和 LeakyReLU(LReLU)组成。

最终，鉴别结果被输入生成网络，为生成网络提供多尺度信息及时空信息反馈，引导生成网络重建视频复杂区域，生成丰富逼真的纹理细节。

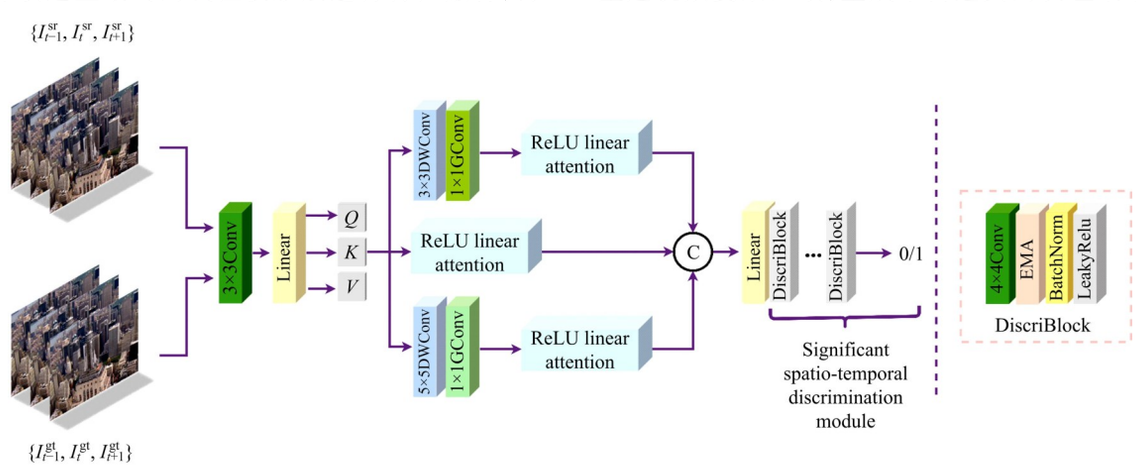

图 4　TRADN 的整体框架示意图

Fig. 4　Illustration of the overall framework of TRADN

1.4　损失函数

本文采用 Charbonnier Loss 作为重建损失，来测量重建样本和真实样本之间的距离，L_{R} 定义如式(17)所示：

$$L_{\mathrm{R}}=\frac{1}{n}\sum_{i=1}^{n}\sqrt{\left\|I_i^{\mathrm{SR}}-I_i^{\mathrm{GT}}\right\|_2^2+\varepsilon} \tag{17}$$

式中：I_i^{SR} 和 I_i^{GT} 分别为在第 i 个时间步长生成的重建样本和真实样本；n 为输入 LR 帧的个数；ε 为一个很小的常数，通常设定为 1×10^{-6}。

对抗性训练是提高 SR 质量的有效手段，对抗损失 L_{adv} 定义如式(18)所示：

$$L_{\mathrm{adv}}=-E_{\mathrm{sr}\sim p_{sr}(\mathrm{sr})}[\log D(I^{\mathrm{GT}})] \tag{18}$$

式中:sr$\sim p_{\mathrm{sr}}$(sr)为重建样本分布。

最后,总损失函数定义如式(19)所示:

$$L=\alpha L_{\mathrm{R}}+\beta L_{\mathrm{adv}} \tag{19}$$

2 实验结果与分析

本节首先介绍数据集(训练集和测试集)、实验配置及评估指标,并将EGTRA与最新的实时或轻量VSR模型进行对比实验。消融实验详细探讨了每个模块的贡献,并对每个模块进行了分析与实验,验证了EGTRA中不同模块的有效性。

2.1 数据集

在实验中,本文使用Vimeo90K[19]数据集作为训练集以训练模型,该数据集包含64 612个用于训练的视频序列,每个序列由7帧组成。该数据集已被广泛认可并用于各种视频相关研究,如VSR和视频插值。实验中对视频帧序列采取4倍下采样,以获得相应LR帧。为获得真实的VSR序列,本文采用标准差为1.5的高斯核对图像进行模糊以实现退化。

为评估在复杂视频场景下,模型对视频纹理细节、遮挡区域的重建效果,在模型测试中,本文采用3个公开的VSR测试集对模型有效性进行评估,具体如下:

Vid4[20]:Vid4测试集被广泛应用于各种视频超分辨率算法中,该测试集由4个不同的运动遮挡场景视频组成,共171帧,其分辨率从480×720至576×704不等。

UDM10[21]:UDM10测试集由10个具有不同动作和场景的高质量视频组成,包含更多纹理和高频细节,共320帧,其分辨率为1280×720。

REDS4[22]:REDS4测试集由4个运动幅度大、场景复杂、遮挡区域多的现实生活场景视频组成,共400帧,其分辨率为1280×720。

这3个测试集的视频场景包含丰富的纹理细节、遮挡区域,而REDS4包含更多大运动幅度下的复杂视频场景,这种视频场景使得VSR模型获取较少的视频信息,进而难以重建视频复杂区域,为研究及解决大运动幅度下纹理细节和遮挡区域多的复杂视频场景问题提供实验基础。本文以这三个测试集为实验基础,设计了EGTRA方法解决实时VSR中难以重建视频复杂区域的问题。

2.2 实验配置及评估指标

本实验环境配置如表1所示,采用Adam优化器对EGTRA进行优化,批次大小设置为8,参数$\beta_1=0.9$,$\beta_2=0.99$,初始学习率为1×10^{-4},在训练过程中使用MultistepLR策略根据步进降低学习率,迭代次数为200 000次,每步进30 000次迭代,学习率降低为原来的1/2,逐步衰减至4×10^{-7}。模型通道数设为64,RRB数量设为10。总损失的平衡系数设置为:$\alpha=0.1$,$\beta=0.01$。约需4 d完成EGTRA的训练。

表1 实验环境配置

Table 1 Experiment environment configuration

名称	环境配置
操作系统	Ubuntu20.04
处理器	Intel(R) Xeon(R) Gold6248 CPU @ 2.50 GHz
显卡	Tesla A100
CUDA版本	11.6
深度学习框架	PyTorch 1.13.0
Python版本	3.9
显存	80 GB

实验中,从3个方面评估模型性能:重建结果的保真度、感知质量以及模型效率。对于重建结果的保真度,采用峰值信噪比(Peak signal to noise ratio,PSNR)和结构相似度(Structural similarity,SSIM)作为评估标准;对于重建结果的感知质量,采用学习感知图像块相似度(Learned perceptual image patch similarity,LPIPS)作为评估标准;对于实时VSR的模型效率,采用模型参数、推理延迟、浮点运算数(FLOPs)以及权衡得分函数score[23]作为评估标准,其中推理延迟表示模型进行VSR所需时间,score用于客观衡量模型在实时下的性能,二者在实时VSR效率评估中极为重要,score定义如式(20)所示:

$$\mathrm{score}=\frac{2^{2\times\mathrm{PSNR}}}{C\times t} \tag{20}$$

式中:C实验中设为250.0的常数,t为推理延迟。

将各对比模型的权衡得分归一化至同一尺度,以便更好地比较。

2.3 实验结果分析

在对比实验中,EGTRA与最新的轻量或实时VSR模型,包括VESPCN[2]、SOFVSR[4]、TecoGAN[13]、FRVSR[3]、EGVSR[9]、SSL[10]、STDO[12]进行比较。其中,VESPCN是一种高效的时空亚像素卷积实时VSR网络,SOFVSR预测高分辨率光流以增强VSR效果,TecoGAN和EGVSR是基于GAN的实时VSR,FRVSR是一

种帧循环 VSR 模型，SSL 通过模型裁剪实现 VSR 模型的轻量化，STDO 是一种利用时空信息过拟合的实时 VSR 模型。

通常，不同的训练集和下采样方式会影响模型的性能，为保证对比的公平性，实验采用相同的训练集以及高斯下采样方式重新训练 VSR 模型。EGTRA 在 3 个测试集上的评估和模型性能对比如表 2 所示，在保真度方面，EGTRA 在 PSNR 和 SSIM 方面显著优于 SSL、STDO、EGVSR 以及 TecoGAN，表明 EGTRA 在 3 个测试集上的重建效果最佳。与 SSL 相比，在 PSNR 和 SSIM 上差距不大，但在感知质量方面，EGTRA 相较于 SSL，在 3 个测试集上的 LPIPS 分别提升了 13.86%、11.05% 和 2.93%，表明 EGTRA 在保持显著重建质量的同时，恢复了更多逼真的纹理细节，更符合人眼的主观感受。图 5～图 7 为不同 VSR 方法的可视化结果。如图 5 所示，在 REDS4 测试集中，面对运动幅度大的视频场景，SSL、STDO、EGVSR、FRVSR 等方法存在严重的失真，缺少清晰的条纹信息，而 EGTRA 通过关注目标区域，获取丰富的多尺度及时空信息，重建样本失真较少，更好地保留了物体细节及感知质量。为更好地说明遮挡区域的重建性能，如图 6、图 7 所示，EGTRA 在目标区域（如标记为遮挡区域的

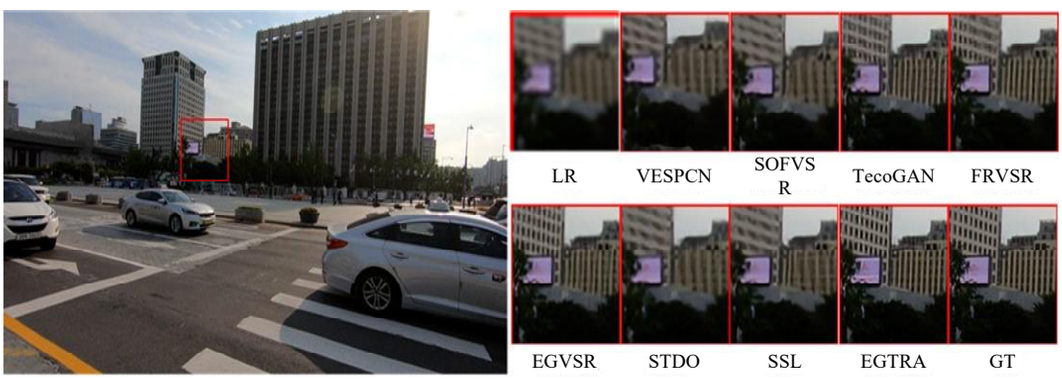

图 5　在 REDS4 测试集上进行 4 倍放大的视觉比较（放大可查看更佳可视化效果）

Fig. 5　Visual comparisons on REDS4 testing set for 4× upscaling (Zoom in to see better visualization)

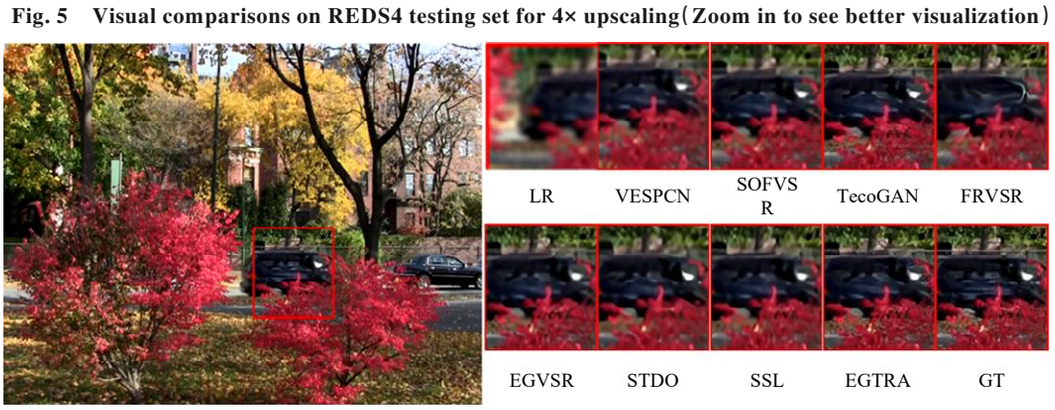

图 6　在 Vid4 测试集上进行 4 倍放大的视觉比较（放大可查看更佳可视化效果）

Fig. 6　Visual comparisons on Vid4 testing set for 4× upscaling (Zoom in to see better visualization)

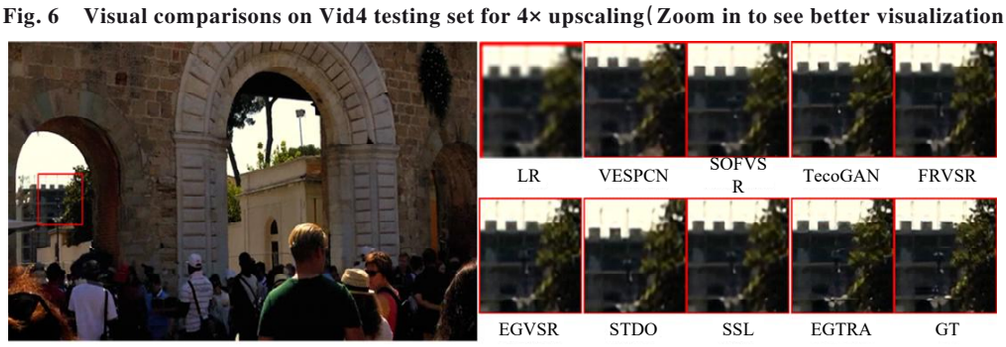

图 7　在 UDM10 测试集上进行 4 倍放大的视觉比较（放大可查看更佳可视化效果）

Fig. 7　Visual comparisons on UDM10 testing set for 4× upscaling (Zoom in to see better visualization)

“灌木与汽车”、“围墙与树叶”)的重建效果优于其他对比模型。可视化结果表明,EGTRA能将复杂视频场景重建为低失真图像,恢复更多复杂区域的条纹细节,重建出高视觉质量的视频。

在模型效率方面,如表3所示,EGTRA在Tesla A-100 GPU上重建REDS4测试集的推理延迟为13.36 ms,相较于对比模型,EGTRA在实时应用中的处理速度最佳,REDS4测试集中视频序列的平均分辨率为180×320,放大因子为4。EGTRA相比于对比模型需要更少的FLOPs,虽然STDO的FLOPs最少,由于对数据的时空拟合,保留的复杂网络结构导致模型参数及推理延迟过度增加。

表2 不同方法在3个VSR测试集上的定量比较

Table 2 Quantitative comparison of different methods on three VSR testing sets

方法	REDS4			Vid4			UDM10		
	PSNR/dB ↑	SSIM ↑	LPIPS ↓	PSNR/dB ↑	SSIM ↑	LPIPS ↓	PSNR/dB ↑	SSIM ↑	LPIPS ↓
VESPCN	22.29	0.594 4	0.334 2	19.70	0.500 2	0.574 7	26.22	0.791 6	0.143 6
SOFVSR	24.31	0.660 2	0.317 2	21.07	0.530 3	0.356 8	27.63	0.816 9	0.131 1
TecoGAN	22.99	0.630 3	0.291 6	24.81	0.770 7	0.200 9	31.97	0.900 9	0.088 8
FRVSR	24.43	0.684 9	0.333 8	25.55	0.798 5	0.280 5	33.95	0.925 6	0.099 5
EGVSR	29.75	0.838 2	0.109 0	25.83	0.796 3	0.137 6	36.96	0.941 1	0.067 6
STDO	29.34	0.844 1	0.189 8	26.30	0.790 2	0.267 6	38.23	0.958 9	0.084 0
SSL	29.81	0.835 6	0.223 7	26.61	0.819 6	0.245 3	38.26	0.957 2	0.079 6
EGTRA	29.89	0.839 9	0.085 1	26.75	0.822 0	0.134 8	38.69	0.966 1	0.050 3

表 3 在REDS4测试集上对模型效率的比较

Table 3 Comparison of model efficiency on the REDS4 testing set

方法	模型参数	FLOPs	推理延迟/ms	实时推理	REDS4评分	REDS4 PSNR
VESPCN	0.879M	96.56G	20.62	√	0.001	22.29
SOFVSR	1.640M	226.12G	75.13	×	0.001	24.31
TecoGAN	2.589M	190.81G	32.10	√	0.001	22.99
FRVSR	2.589M	190.81G	32.09	√	0.014	24.43
EGVSR	2.681M	102.89G	14.25	√	50.812	29.75
STDO	4.843M	2.76G	35.71	√	11.485	29.34
SSL	0.589M	63.9G	18.25	√	43.117	29.81
EGTRA	2.177M	60.69G	13.36	√	65.806	29.89

综上所述,与对比模型相比,EGTRA具有极低的推理延迟,在重建质量方面具有显著的性能竞争力。同时,EGTRA在REDS4测试集上获得更高的权衡得分,达到了65.806,比排名第二的EGVSR高出14.994,有效证明了该模型在实时应用中优于其他对比方法。

2.4 消融实验分析

为验证本文所提模块的有效性,首先评估SCA模块的有效性,随后评估RRB的有效性。此外,对多尺度支路、EGRN、RLA及SSTDM模块的有效性进行全面消融实验。

为验证SCA模块的有效性,进行消融实验,其结果如表4所示,SCA-o表示未使用SCA模块的模型,SCA-w表示使用SCA模块的模型,可以看出,移除SCA模块后,模型参数、FLOPs、推理延迟几乎不变,与未使用SCA的模型相比,完整模型在REDS4、Vid4、UDM10测试集上的PSNR性能分别提升0.12、0.07、0.29 dB,结果表明,网络性能可通过少量额外的FLOPs和参数得到改善,消融实验结果证明了SCA模块的有效性。

针对RRB模块的有效性,本文亦进行消融实验,RRB-o表示未使用RRB的模型,RRB-w表示使用RRB的模型,其结果如表5所示,可以观察到,完整模型在保持定量性能的同时,模型参数、FLOPs、推理延迟等效率指标得到显著下降,其中推理延迟下降了20%,进一步提升了EGTRA的推理效率。

如表6所示,本文对目标区域关注鉴别网络

的多尺度支路进行消融评估，Model1 表示移除了 3×3 与 5×5 支路的模型，model 2 表示模型只保留 3×3 支路，model 3 表示模型只保留 5×5 支路，model 4 表示模型拥有全部支路。根据结果分析，在没有支路的情况下，Model 1 在测试集上的指标都是最低的，表明模型缺乏对目标区域的特征提取能力。相比之下，同时拥有两条支路的模型 Model 4 可以有效提取视频多尺度信息及区域多尺度信息。PSNR 在 3 个测试集分别提升 0.3、0.21、0.21 dB，LPIPS 分别提升 2.25%、1.00%、2.92%，进一步说明了 RLA 的必要性，此外，Model 2 与 Model 3 的各项定量指标均优于 Model1，证明各支路的有效性。

如表 7 所示，本文对各个模块的有效性进行了消融实验，SSTDM-o 表示模型缺少 SSTDM，RLA-o 表示模型缺少 RLA，EGRN-o 表示模型缺少 EGRN，Model 即为 EGTRA。Model 与 EGRN-o 相比，在 3 个测试集上的 PSNR 分别提升 0.48、0.46、0.49 dB，SSIM 分别提升 0.91%、0.78%、0.56%，说明 EGRN 模块能够减少时间聚合误差带来的影响，提升模型重建质量。Model 与 RLA-o 相比，在 3 测试集上的 LPIPS 分别提升 1.95%、1.00%、2.92%，表明 RLA 模块可以显著提升感知质量。Model 与 SSTDM-o 相比，说明 SSTDM 对各指标都有提升。显著的性能增益表明，EGRN、RLA 和 SSTDM 的组合能够在多个指标上提升 VSR 重建效果。消融实验验证了所设计模块在不同复杂视频场景中的有效性。

表 4　三个 VSR 测试集上 SCA 模块的有效性验证

Table 4　Validation of SCA module on three VSR testing sets

方法	模型参数	FLOPs	推理延迟/ms	REDS4			UDM10			Vid4		
				PSNR/dB ↑	SSIM ↑	LPIPS ↓	PSNR/dB ↑	SSIM ↑	LPIPS ↓	PSNR/dB ↑	SSIM ↑	LPIPS ↓
SCA-o	2.175M	60.69G	13.32	29.77	0.835 0	0.085 3	26.68	0.8184	0.135 1	38.40	0.963 5	0.051 7
SCA-w	2.177M	60.69G	13.36	29.89	0.839 9	0.085 1	26.75	0.8220	0.134 8	38.69	0.966 1	0.050 3

表 5　三个 VSR 测试集上 RRB 模块的有效性验证

Table 5　Validation of RRB module on three VSR testing sets

方法	模型参数	FLOPs	推理延迟/ms	REDS4			UDM10			Vid4		
				PSNR/dB ↑	SSIM ↑	LPIPS ↓	PSNR/dB ↑	SSIM ↑	LPIPS ↓	PSNR/dB ↑	SSIM ↑	LPIPS ↓
RRB-o	2.861M	102.89G	16.68	29.94	0.841 2	0.085 0	26.78	0.824 5	0.134 8	38.76	0.966 4	0.050 3
RRB-w	2.177M	60.69G	13.36	29.89	0.839 9	0.085 1	26.75	0.822 0	0.134 8	38.69	0.966 1	0.050 3

表 6　三个 VSR 测试集上多尺度支路的有效性验证

Table 6　Validation of multiscale branches on three VSR testing sets

方法	3×3 支路	5×5 支路	REDS4			UDM10			Vid4		
			PSNR/dB ↑	SSIM ↑	LPIPS ↓	PSNR/dB ↑	SSIM ↑	LPIPS ↓	PSNR/dB ↑	SSIM ↑	LPIPS ↓
Model 1			29.59	0.832 2	0.107 6	26.54	0.816 2	0.144 8	38.48	0.962 3	0.079 5
Model 2	√		29.68	0.834 1	0.095 0	26.61	0.816 3	0.142 9	38.54	0.963 3	0.058 3
Model 3		√	29.64	0.834 8	0.097 4	26.66	0.818 9	0.138 0	38.56	0.963 8	0.069 2
Model 4	√	√	29.89	0.839 9	0.085 1	26.75	0.822 0	0.134 8	38.69	0.966 1	0.050 3

表 7　三个 VSR 测试集上 EGTRA 各模块的有效性验证

Table 7　Validation of EGTRA modules on three VSR testing sets

方法	EGRN	RLA	SSTDM	REDS4			UDM10			Vid4		
				PSNR/dB ↑	SSIM ↑	LPIPS ↓	PSNR/dB ↑	SSIM ↑	LPIPS ↓	PSNR/dB ↑	SSIM ↑	LPIPS ↓
SSTDM－o	√	√		29.86	0.838 1	0.089 3	26.71	0.819 8	0.143 9	38.59	0.964 9	0.064 0
RLA－o	√		√	29.59	0.832 2	0.107 6	26.54	0.816 2	0.144 8	38.48	0.964 1	0.079 5
EGRN－o		√	√	29.41	0.830 8	0.085 2	26.29	0.814 2	0.135 1	38.20	0.960 5	0.050 8
Model	√	√	√	29.89	0.839 9	0.085 1	26.75	0.822 0	0.134 8	38.69	0.966 1	0.050 3

3 结束语

本文提出了一种基于高效门控和目标区域关注的实时视频超分辨率模型,用于解决在面临运动幅度较大的复杂视频场景,实时视频超分辨率算法难以重建纹理细节与遮挡区域的问题。基于生成对抗网络,本文使用高效门控重建网络作为生成网络,可有效缓解时间聚合误差的影响,门控机制自适应地提取复杂区域特征,重参数化残差块保证了高效的模型性能。在此基础上,本文进一步提出了目标区域关注鉴别器网络,为生成网络提供多尺度及时空信息反馈,引导生成网络重建复杂区域纹理细节。与其他先进方法相比,在保持显著保真度的同时,与保真度排名第二的SSL相比,本文模型在3个测试集上的LPIPS指标分别提升了13.86%、11.05%、2.93%,并以13.36 ms的推理延迟处理视频序列,同时获得了65.208的实时得分,在重建质量与效率之间实现了最佳权衡。未来工作中,将探索在保证高效处理的同时,结合神经渲染技术,进一步提升其视频超分辨质量。

参考文献:

[1] Rota Claudio, Buzzelli Marco, Bianco Simone, et al. Video restoration based on deep learning: a comprehensive survey[J]. Artificial Intelligence Review, 2023, 56(6): 5317-5364.

[2] Caballero J, Ledig C, Aitken A P, et al. Real-time video super-resolution with spatio-temporal networks and motion compensation[C]//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2017: 4778-4787.

[3] Vemulapalli R, Brown M, Mehdi S M. Frame-recurrent video super-resolution[C]//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Salt Lake City, UT, USA, 2018: 6626-6634.

[4] Wang L, Guo Y, Liu L, et al. Deep video super-resolution using HR optical flow estimation[J]. IEEE Transactions on Image Processing, 2020, 29: 4323-4336.

[5] Chan K C K, Wang X, Yu K, et al. Basicvsr: the search for essential components in video super-resolution and beyond[C]//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, TN, USA, 2021: 4947-4956.

[6] Chan K C K, Zhou S, Xu X, et al. Basicvsr++: improving video super-resolution with enhanced propagation and alignment[C]//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, New Orleans, LA, USA, 2022: 5972-5981.

[7] Ouyang Ning, Ou Zhi-shan, Lin Le-ping. Video super-resolution network with gated high-low resolution frames [J]. Applied Sciences, 2023, 13(14): 1-16.

[8] Zhou X, Zhang L, Zhao X, et al. Video super-resolution transformer with masked inter & intra-frame attention[C]//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2024: 25399-25408.

[9] Cao Y, Wang C, Song C, et al. Real-time super-resolution system of 4k-video based on deep learning [C]//2021 IEEE 32nd International Conference on Application-Specific Systems, Architectures and Processors(ASAP), NJ, USA, 2021: 69-76

[10] Xia Bin, He Jing-wen, Zhang Yu-lun, et al. Structured sparsity learning for efficient video super-resolution[C]//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. Vancouver, Canada. 2023: 22638-22647

[11] Xiao Jun, Jiang Xin-yang, Zheng Ning-xin, et al. Online video super-resolution with convolutional kernel bypass grafts[J]. IEEE Transactions on Multimedia, 2023, 25: 8972-8987.

[12] Li Gen, Ji Jie, Qin Ming-hai, et al. Towards high-quality and efficient video super-resolution via spatial-temporal data overfitting[C]//2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). IEEE, Vancouver, Canada. 2023: 10259-10269.

[13] Chu Meng-yu, Xie You, Leal-Taixé Laura, et al. Temporally coherent gans for video super-resolution (tecogan)[J]. arXiv Preprint, 2018, 1(2): 3.

[14] Chen Rui, Mu Yang, Zhang Yan. High-order relational generative adversarial network for video super-resolution [J]. Pattern Recognition, 2024, 146: 110059.

[15] Chen L, Chu X, Zhang X, et al. Simple baselines for image restoration[C]//European Conference on Computer Vision. Cham: Springer, 2022: 17-33.

[16] Ding X, Zhang X, Han J, et al. Diverse branch block: Building a convolution as an inception-like unit [C]//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, TN, USA, 2021: 10881-10890,.

[17] Cai H, Li J, Gan C, et al. Efficientvit: Lightweight multi-scale attention for high-resolution dense prediction[C]//Proceedings of the IEEE/CVF International Conference on Computer Vision. Paris, France. 2023: 17302-17313.

[18] Ouyang D, He S, Zhan J, et al. Efficient multi-scale attention module with cross-spatial learning[C]// ICASSP 2023-2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). ArXiv, 2023, abs/2305.13563.

[19] Xue Tian-fan, Chen Bai-an, Wu Jia-jun, et al. Video enhancement with task-oriented flow[J]. International Journal of Computer Vision, 2019, 127: 1106-1125.

[20] Liu Ce, Sun De-qing. On Bayesian adaptive video super resolution[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2013, 36(2): 346-360.

[21] Yi P, Wang Z, Jiang K, et al. Progressive fusion video super-resolution network via exploiting non-local spatio-temporal correlations[C]//Proceedings of the IEEE/CVF International Conference on Computer Vision, Seoul, Korea (South), 2019: 3106-3115.

[22] Nah S , Baik S, Hong S, et al. Ntire 2019 challenge on video deblurring and super-resolution: Dataset and study[C]//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops, Long B, CA, USA, 2019, 1996-2005.

[23] Ignatov A, Romero A, Kim H, et al. Real-time video super-resolution on smartphones with deep learning, Mobile AI2021 challenge: Report[C]//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Nashville, TN, USA, 2021: 2535-2544.